Modulation-Frequency Encoding of Speech with Applications to Neural Speech Recognizers

M. Paping, H. W. Strube, T. Gramss[†]

Drittes Physikalisches Institut, University of Göttingen
Bürgerstrasse 42–44; D-37073 Göttingen; Germany
Tel: +49(551)397731; Fax: +49(551)397720
e-mail: paping@up3spr2.gwdg.de

Most of the present speech recognition systems make use of a signal representation based on parameter frames equidistant in time. We propose a new method of signal representation which is motivated by linguistic, psychoacoustic and neurobiological facts.

The basic idea of our approach is a nonlinear segmentation of the time signal. Starting with a frame-based Bark spectrogram (stepsize 10 ms, 40 ms Hamming window) the segment boundaries are defined by the local minima of the smoothed modified loudness. We get the modified loudness by summing up the weighted frequency components of each time frame of the Bark spectrogram. The length of the resulting segments range from demisyllables to syllables. Seen from a linguistic point of view there are several reasons for choosing such large segments. One advantage may be that coarticulation does not affect syllables so much as it affects phonemes.

For each of these segments a feature vector is calculated. Motivated by recent results in modern brain research we perform another Fourier transformation on each frequency channel of the Bark spectrogram. In doing so we get a signal representation in the three-dimensional modulation frequency space. This kind of sound encoding is found in the auditory midbrain of several mammals and birds. Since the temporal structure of a signal is encoded in a redundant way now, it is possible to integrate in time over the segments defined above. As shown in previous papers this can be done without causing too much loss of information.

Now a speech signal is represented by a sequence of feature vectors, whose number of components is constant and thus independent of the length of the respective speech segment. This implies a local time normalization. The average length of the segments corresponds well to the time window of about 200 ms up to which an acoustical event can be integrated in the human auditory system.

Before starting recognition experiments we focused on the segmentation problem. The aim was to get a phonetically consistent segmentation: different versions of an utterance should be separated at phonetically equivalent points. For this task a special optimization procedure has been developed. The free parameters were the weighting window, used to calculate the modified loudness, and the degree of smoothing. We found that best results were reached with a trapezoidal window.

An artificial neural network is then used to perform isolated word recognition. Because of the feature vectors' high dimension a two-layer perceptron is sufficient for that purpose, rendering computation very efficient. In the training mode the simple δ-rule is used to change the synaptic plasticities in such a way that each feature vector of a given word is mapped onto the output cell representing this word. In the test mode each feature vector of a given utterance is applied to the neural network, resulting in a sequence of probability distributions from the output layer. From this sequence the recognized word can be estimated by multiplying all probabilities of each word. The winner cell is then determined by the product with maximal activity.

The capability of the recognition system was tested on the SPINA data base which consists of 62 isolated German words (five male and five female speakers). The speaker dependent recognition rates ranged from 90.7% to 99.4% (average 94.7%).

This new approach is suitable for being extended to processing continuous speech.

[†] current address: Physics of Computation Lab., Californian Institute of Technology, Pasadena, USA

S. Gielen et al. (eds.), *ICANN '93*